ANALYSIS OF PROCTOR MARKING ACCURACY IN A COMPUTER-AIDED PERSONALIZED SYSTEM OF INSTRUCTION COURSE

TOBY L. MARTIN, JOSEPH J. PEAR, AND GARRY L. MARTIN

UNIVERSITY OF MANITOBA

In a computer-aided version of Keller's personalized system of instruction (CAPSI), students within a course were assigned by a computer to be proctors for tests. Archived data from a CAPSI-taught behavior modification course were analyzed to assess proctor accuracy in marking answers as correct or incorrect. Overall accuracy was increased by having each test marked independently by two proctors, and was higher on incorrect answers when the degree of incorrectness was larger.

DESCRIPTORS: instructional design, personalized system of instruction, online teaching

Recent years have seen growth in computer-mediated education at the university level, but research on the effectiveness of such course procedures has not kept pace. In this paper we analyze proctor accuracy in a computer-mediated course procedure called computer-aided personalized system of instruction (CAPSI) in use at the University of Manitoba (Pear & Crone-Todd, 1999). CAPSI shares many features with the personalized system of instruction (PSI) first described by Keller (1968). Like PSI, CAPSI provides students with clear study questions that are based on the content of written materials rather than the content of lectures. The study questions are grouped into small study units; students must pass a test on each unit before attempting a test on the next unit. Tests are attempted at the student's own pace, and as often as time permits. If a student is unsuccessful on a unit test, he or she may attempt a new test on that unit.

Central to both CAPSI and PSI is the use of student evaluators, called proctors, who mark unit tests and provide feedback. Proctors are a major source of feedback, so it is important that they provide the most accurate feedback possible. The present study is the first to assess the accuracy of the feedback given by CAPSI proctors.

METHOD

Participants and Database

The participants were 33 students who, in the fall of 1996, completed an undergraduate psychology course at the University of Manitoba. Study questions from the first 15 chapters of the course textbook (G. L. Martin & Pear, 1996) were grouped into 10 study units. As each student worked through these units, the CAPSI program (a) randomly generated and electronically delivered to a student, upon his or her request, short essay-type tests based on the study questions from the appropriate unit; (b) assigned markers to each completed unit test (a marker is anyone whom the program selected to evaluate a test, including the instructor and teaching assistant; a proctor is a marker who was another student in the course); (c) electroni-

This research was part of a thesis submitted by the first author in partial fulfillment of the requirements for a master's degree at the University of Manitoba. It was supported in part by a grant from the Social Sciences and Humanities Research Council of Canada to J. J. Pear. We are grateful to Darlene Crone-Todd for comments on an earlier draft of the article.

Address correspondence to T. L. Martin or J. J. Pear at the Department of Psychology, University of Manitoba, Winnipeg R2T 2N2, Canada (e-mail: ummarti7@cc.umanitoba.ca or pear@cc.umanitoba.ca).

cally delivered the completed test to the selected markers; and (d) electronically returned the test to the student with written feedback from the markers. The information recorded in the CAPSI files constituted the database of the study.

A student became eligible to be a proctor for a particular unit when he or she passed a test on that unit. Whenever possible, two proctors were assigned to mark each unit test. If there were not two students eligible and available to be proctors for a unit test, then either the instructor or the teaching assistant was assigned to mark that test. A proctor's primary responsibility was to determine whether the student passed the test within 24 hr of being assigned to mark it. Proctors were instructed that assigning a pass was appropriate only when all answers demonstrated mastery of the relevant concepts. Both proctors had to assign a pass for the student to receive a mark of pass on the test.

For this study, a sample of 101 unit tests (19.3% of the total 523 unit tests from the course) was selected by identifying tests containing questions that were also randomly selected for inclusion on a subsequent unit test, the midterm exam, or the final exam. These criteria permitted additional analyses of feedback and are described in greater detail by T. L. Martin, Pear, and Martin (in press).

Two assessors independently determined the correctness of the 302 answers in the sample using the same criteria as the proctors. The first assessor had previously passed the course with high marks and repeatedly served as a teaching assistant for the course. The second assessor was an individual with many years of experience teaching the course. An interobserver reliability score was calculated as the number of agreements divided by the number of agreements plus disagreements multiplied by 100%. This yielded a reliability score of 83% (taken over all 302 answers). For each disagreement, the merits and deficiencies of the answer were discussed in order to reach agreement. Agreement could not be reached on three answers, which were discarded to produce a final sample of 299 answers.

The marking determinations made by the proctors were compared to the standard obtained by the agreement and consensus of the two experts to determine the accuracy of the 559 instances of proctoring (IOPs) in the sample. One proctor marking one answer equaled one IOP; hence a single answer usually corresponded to two IOPs. In addition, IOPs on answers on which experts initially disagreed but for which consensus was obtained following discussion were considered to be IOPs on "difficult" answers for the purposes of data analysis.

RESULTS AND DISCUSSION

Part A of Figure 1 shows how accurately the proctors marked all of the answers relative to the consensus of the experts. In 187 IOPs on incorrect answers, proctors erroneously marked an answer as correct (i.e., produced a false negative) in 125 cases (66.8%). Proctors were much more accurate when marking answers that the experts marked as correct. In 372 IOPs on correct answers, proctors erroneously marked an answer as incorrect (i.e., produced a false positive) in 25 cases (6.7%). Proctors made errors in 26.8% of all IOPs. Of all proctor errors, 83% were false negatives. Of all incorrect answers, 53.6% were marked as incorrect by at least one proctor; thus, having two proctors mark each test reduced the percentage of false negatives from 66.8% to 46.4%.

Overall, these results suggest that the proctors were biased toward marking answers as correct, because the proctors agreed with the consensus of the experts on the majority of correct answers but disagreed with the experts on the majority of incorrect answers. This bias may have reflected skill def-

A. All Instances of Proctoring (IOPs)

		Expert Consensus on Answers	
		Incorrect	Correct
Proctors' Determinations	Incorrect	62 True Positives	25 False Positives
	Correct	125 False Negatives	347 True Negatives

B. Instances of Proctoring Difficult Answers

		Expert Consensus on Answers	
		Incorrect	Correct
Proctors' Determinations	Incorrect	6 True Positives	0 False Positives
	Correct	32 False Negatives	10 True Negatives

Figure 1. Breakdown of proctor accuracy.

icits (e.g., the proctors had difficulty detecting errors), motivational factors (e.g., it required less time and effort to give positive feedback than negative feedback), or both. However, the proctors' performance on difficult answers (i.e., ones on which the experts initially disagreed) suggests that motivational factors did not account for all false negatives.

Part B of Figure 1 shows how accurately the proctors performed on difficult answers. In 38 IOPs on incorrect difficult answers, proctors erroneously marked an answer as correct (i.e., produced a false negative) in 32 cases (84.2%). In 10 IOPs on correct difficult answers, proctors produced no false positives.

Individual error rates (total errors as a percentage of total IOPs for a given proctor) ranged from 0% to 50%. Proctors differed widely in their detection of incorrect answers. For example, the proctor with the most instances of proctoring (50 IOPs) encountered 12 incorrect answers and marked them all as correct, whereas the third busiest proctor (37 IOPs) encountered 16 incorrect answers and marked only two of them as correct.

Although limited by its descriptive nature, the present study makes a unique contribution by assessing feedback accuracy in a CAPSI-taught course. There is a paucity of data on the accuracy and effectiveness of feedback in postsecondary courses. The present study, in conjunction with that by T. L. Martin et al. (in press), helps to fill this gap and provides a methodology that could lead to enhanced feedback in various course procedures, especially those involving online teaching.

REFERENCES

Keller, F. S. (1968). Good-bye, teacher . . . *Journal of Applied Behavior Analysis, 1,* 79–89.

Martin, G. L., & Pear, J. J. (1996). *Behavior modification: What it is and how to do it* (6th ed.). Upper Saddle River, NJ: Prentice Hall.

Martin, T. L., Pear, J. J., & Martin, G. L. (in press). Proctor feedback and its effectiveness in a computer-aided personalized system of instruction course. *Journal of Applied Behavior Analysis.*

Pear, J. J., & Crone-Todd, D. E. (1999). Personalized system of instruction in cyberspace. *Journal of Applied Behavior Analysis, 32,* 205–209.

Received May 8, 2001
Final acceptance May 20, 2002
Action Editor, Craig H. Kennedy